Understanding Reward Hacking in Text-to-Image Reinforcement Learning

Yunqi Hong[1]; Kuei-Chun Kao[1]; Hengguang Zhou[1]; Cho-Jui Hsieh[1,2]
[1]University of California, Los Angeles, [2]Arena

CVPR JUNE 3-7, 2026 DENVER COLORADO

Abstract

RL is widely used to improve text-to-image generation, but imperfect reward models can lead to reward hacking, where images receive high reward scores despite poor visual quality[1].

We systematically analyze reward hacking across aesthetic, preference, and prompt-alignment rewards, and find a common failure mode: artifact-prone image generation.

To address this issue , we propose ArtifactReward, a lightweight reward model trained on a small curated dataset to penalize visual artifacts and improves image quality.

Main Contributions

- ❑ Measurement Framework
  - We build a framework to observe and compare T2I reward hacking behaviors across multiple reward models.
- ❑ Reward → Behavior
  - We show how different reward properties largely determine image generation model behaviors.
- ❑ A Common Failure Mode
  - Across multiple reward settings, RL frequently produces structural artifacts rather than semantic improvements.
- ❑ Mechanism-Driven Mitigation
  - We explore mitigation through artifact-aware reward augmentation.

Our code is available at: https://github.com/yq-hong/ArtifactReward

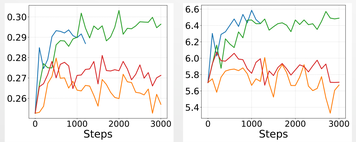
Reward Hacking

➢ We analyze how different rewards lead to reward hacking in T2I RL across diverse settings.

- Aesthetic / Human preference: HPS[2]
- Text-Image Alignment: GDino[3], ORM[4]
- Ensemble multiple rewards

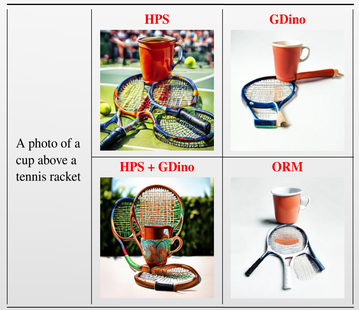
❑ Reward-specific improvement and cross-metric degradation.

- RL substantially improves the optimized reward metric but yields limited gains on other metrics, indicating narrow reward optimization rather than genuine improvement in image quality.



(a) HPS - Object (b) Aesthetic Score - Object

(e) GDino - Spatial (f) ORM - Color



Training Reward: Blue: HPS Orange: GDino Red: ORM Green: HPS + GDino

Reward-driven Bias

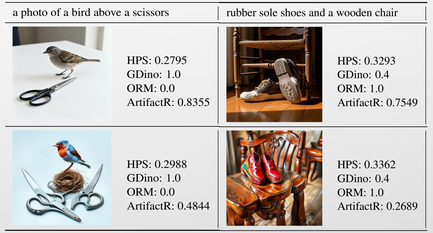
❑ Different rewards induce their own visual biases, often improving reward scores at the expense of overall image quality.

Reward Type	Optimization Tendency	Side Effect
HPS	Vibrant colors, dramatic lighting, rich backgrounds	Over-stylization, over-saturation, weaker spatial fidelity
GDino / ORM	Simple scenes, object-centered compositions	Lower diversity, flatter and less realistic images

A Shared Blind Spot Across Rewards

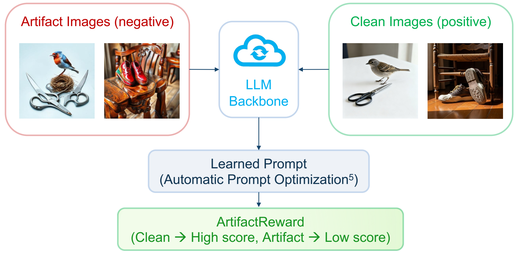
❑ Common failure mode identified across all settings: Generation of artifact-prone images.

- Geometric distortions (e.g., malformed objects)
- Duplication / fragmentation (e.g., extra heads, broken boundaries)
- Entity blending (e.g., merged objects or identities)



Our Solution: ArtifactReward

❑ We propose ArtifactReward, a lightweight and adaptive reward to augment existing reward and suppress reward hacking.



Consistent Improvements

❑ Adding our ArtifactReward consistently improves realism and reduced reward hacking across multiple T2I RL setups.

- Better Visual Realism (Fewer artifacts, more plausible images)
- Stronger Consistency (Higher alignment with prompts)
- Robust Across Benchmarks (Improved result on WISE and LLM4LLM)

A teacher in a white blouse stands at the blackboard, her curly brown hair tied back in a ponytail.

(a) HPS (b) HPS + ArtifactR (c) GDino (d) GDino + ArtifactR

(e) HPS + GDino (f) HPS + GDino + ArtifactR (g) ORM (h) ORM + ArtifactR

Contact

Yunqi Hong
UCLA
Email: yunqihong@ucla.edu
Website: https://yq-hong.github.io/

References

1. Skalse, Joar, et al. "Defining and characterizing reward gaming." Advances in Neural Information Processing Systems 35 (2022): 9460-9471.
2. Wu, Xiaoshi, et al. "Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis." arXiv preprint arXiv:2306.09341 (2023).
3. Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." European conference on computer vision. Cham: Springer Nature Switzerland, 2024.
4. Guo, Ziyu, et al. "Can We Generate Images with CoT? Let's Verify and Reinforce Image Generation Step by Step." arXiv preprint arXiv:2501.13926 (2025).
5. Pryzant, Reid, et al. "Automatic prompt optimization with "gradient descent" and beam search." Proceedings of the 2023 conference on empirical methods in natural language processing. 2023.